

Raport - projekt końcowy z zajęć "Analiza i wizualizacja danych"

Aleksandra Ulaszewska

Grudzień 2025

Spis treści

1	Dane	3
2	Metody	5
2.1	Problem pierwszy - test istotności różnic dla danych niezależnych	5
2.2	Problem drugi - test istotności różnic dla danych niezależnych	5
2.3	Problem trzeci - test siły związku	5
2.4	Problem czwarty - test siły związku	6
2.5	Problem piąty - test istotności różnic dla danych niezależnych	6
3	Wyniki	7
3.1	Problem pierwszy - test istotności różnic dla danych niezależnych	7
3.2	Problem drugi - test istotności różnic dla danych niezależnych	9
3.3	Problem trzeci - test siły związku	11
3.4	Problem czwarty - test siły związku	13
3.5	Problem piąty - test istotności różnic dla danych niezależnych	15

1 Dane

Użyte przeze mnie dane dotyczą higieny snu oraz trybu życia prowadzonego przez osoby, których dane zostały zebrane. Zbiór danych pobrałam ze strony Kaggle.com (<https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset>) i został on stworzony przez panią Laksikę Tharmalingam w celach naukowych. Każdy z wierszy odnosi się do pojedynczej osoby, która udzieliła informacji na swój temat. Łącznie zebrano odpowiedzi od 374 osób. Zbiór danych zawiera 13 kolumn (zmienne) i są to kolejno:

1. numer identyfikacyjny osoby (numer ID) - zmienna ilościowa
2. płeć (kobieta, mężczyzna) - zmienna nominalna
3. wiek - zmienna ilościowa
4. wykonywany zawód
5. ilość snu (ilość snu na dzień przedstawiona w godzinach) - zmienna ilościowa
6. jakość snu (oceniana subiektywnie przez osoby podające swoje dane w skali od 1 do 10) - zmienna porządkowa
7. poziom aktywności fizycznej (minuty wykonanych ćwiczeń w danym dniu) - zmienna ilościowa
8. poziom stresu (oceniany subiektywnie przez osoby podające swoje dane w skali od 1 do 10) - zmienna porządkowa
9. BMI (kategoria BMI (w normie, powyżej normy) danej osoby) - zmienna nominalna
10. ciśnienie krwi (ciśnienie skurczowe i rozkurczowe danej osoby) - zmienna ilościowa
11. tętno (tętno spoczynkowe osoby podającej dane w uderzeniach na minutę) - zmienna ilościowa
12. dzienna ilość kroków - zmienna ilościowa
13. zaburzenia snu (brak, insomnia, bezdech senny) - zmienna nominalna

W moim raporcie użyłam zmiennych:

- płeć

- poziom stresu
- ilość snu
- poziom aktywności fizycznej
- jakość snu
- BMI

Zbiór oryginalnie nie posiadał znaczącej ilości braków danych, jednak mimo tego zdecydowałam się na dodatkową ich filtrację w celu zlokalizowania, a następnie usunięcia możliwych braków. Z racji, że dane oryginalnie występują w języku angielskim zmieniałam nazwy kolumn i niektórych wartości na język polski, aby lepiej przedstawić wyniki w raporcie (np. sleep disorder = zaburzenia snu, sleep apnea = bezdech senny). W kolumnie "płeć" nazwy wartości zostały skrócone do pierwszej litery (male = M, female = K) aby zachować większą przejrzystość danych. Dodatkowo zmienna "BMI" została zdychotomizowana i zamiast trzech wartości (normal, overweight, obese) zawiera dwie (w normie (normal) i powyżej normy (overweight, obese)), aby lepiej zbadać jedną z postawionych hipotez. Następnie zmieniałam typ kolumny "ilość snu" z float na int. Zmiana ta została wdrożona, aby ujednolicić typy zmiennych oraz uprościć analizę. Nie użyto polskich znaków w zapisie kolumn i wartości zmiennych.

2 Metody

Utworzono i przeanalizowano 5 problemów badawczych na podstawie wybranego zbioru danych. Przyjęty poziom ufności jest równy 0.05α .

2.1 Problem pierwszy - test istotności różnic dla danych niezależnych

Pytanie badawcze: Czy kobiety i mężczyźni różnią się poziomem stresu?

Hipoteza badawcza: Kobiety i mężczyźni różnią się poziomem stresu.

Użyta metoda: test Manna-Whitneya

2.2 Problem drugi - test istotności różnic dla danych niezależnych

Pytanie badawcze: Czy kobiety i mężczyźni różnią się ilością snu?

Hipoteza badawcza: Kobiety i mężczyźni różnią się ilością snu.

Test wstępny sprawdzający normalność rozkładu: test Shapiro-Wilka

Wynik testu wstępnego: Wynik jest istotny statystycznie na poziomie istotności 0,05 ($W = 0,81$, $p \leq 0,001$), co oznacza brak rozkładu normalnego dla wskaźnika ilości snu w grupach o różnym wskaźniku płci.

Użyta metoda: test Manna-Whitneya

2.3 Problem trzeci - test siły związku

Pytanie badawcze: Czy między wskaźnikami ilości snu a poziomem aktywności fizycznej istnieje związek?

Hipoteza badawcza: Korelacja między wskaźnikiem ilości snu a wskaźnikiem poziomu aktywności fizycznej jest różna od 0.

Test wstępny sprawdzający założenia korelacji Pearsona o normalności rozkładu: test Henze-Zirklera

Wynik testu wstępnego: Wynik jest istotny statystycznie na poziomie istotności 0,05 ($HZ = 73,71$, $p \leq 0,001$), co oznacza niespełnienie założenia normalności w teście korelacji Pearsona dla wskaźnika ilości snu i wskaźnika poziomu aktywności fizycznej.

Użyta metoda: współczynnik korelacji Spearmana

2.4 Problem czwarty - test siły związku

Pytanie badawcze: Czy między wskaźnikami poziomu stresu i jakości snu istnieje związek?

Hipoteza badawcza: Korelacja między wskaźnikiem poziomu stresu a wskaźnikiem jakości snu jest różna od 0.

Użyta metoda: współczynnik korelacji Spearmana

2.5 Problem piąty - test istotności różnic dla danych niezależnych

Pytanie badawcze: Czy osoby o odmiennym BMI różnią się poziomem aktywności fizycznej?

Hipoteza badawcza: Osoby o odmiennym BMI różnią się poziomem aktywności fizycznej.

Test wstępny sprawdzający normalność rozkładu: test Shapiro-Wilka

Wynik testu wstępnego: Wynik jest istotny statystycznie na poziomie istotności 0,05 ($W = 0,90$, $p \leq 0,001$), co oznacza brak rozkładu normalnego dla wskaźnika poziomu aktywności fizycznej w grupach o różnym wskaźniku BMI.

Użyta metoda: test Manna-Whitneya

3 Wyniki

3.1 Problem pierwszy - test istotności różnic dla danych niezależnych

Wykres:



Rysunek 1: Wykres pierwszy - rozkład współczynnika poziomu stresu względem płci

Hipotezy testowe:

H0: Mediana poziomu stresu u kobiet jest równa medianie poziomu stresu u mężczyzn.

H1: Mediana poziomu stresu u kobiet nie jest równa medianie poziomu stresu u mężczyzn.

Tabela 1: Porównanie poziomu stresu między płciami

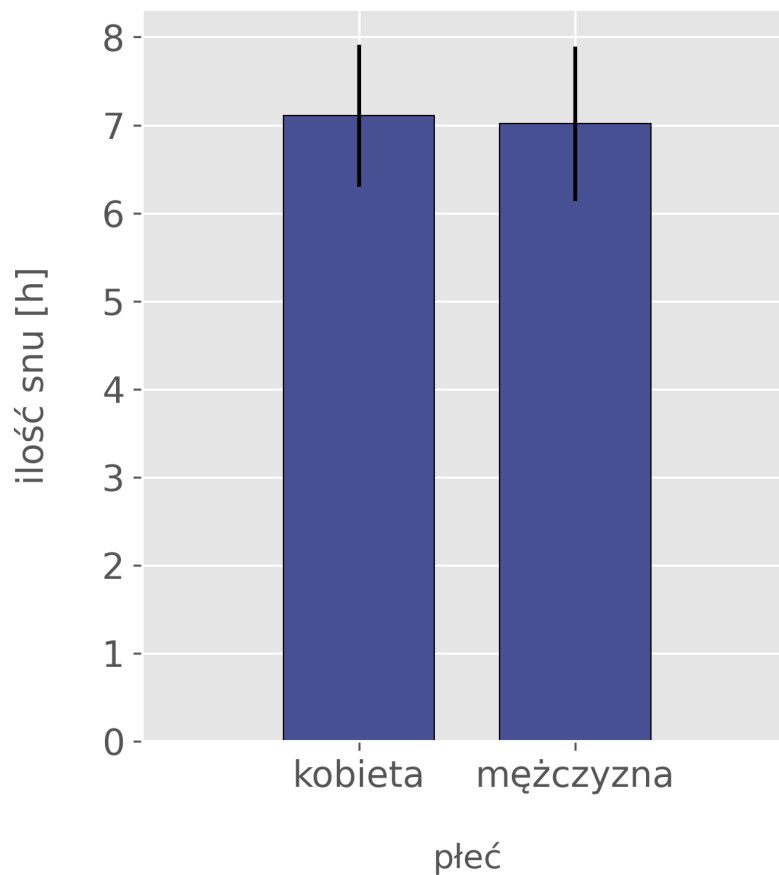
płeć	poziom stresu			test Manna–Whitneya		
	n	med	IQR	U	p^*	CL
kobieta	185	2	3	9034	$\leq 0,001$	0,26
mężczyzna	189	4	2			

* – $p \leq 0,05$

Wniosek: Statystyka testowa oznaczona została jako U , med odpowiada medianie obu analizowanych grup, a IQR oznacza rozstęp międzykwartylowy. Zgodnie z poziomem istotności $p = 0,05$ odrzucamy hipotezę zerową (H_0), mówiącej o równości mediany współczynnika stresu w obu grupach. Kobiety ($med = 2$, $IQR = 3$) i mężczyźni ($med = 4$, $IQR = 2$) różnią się poziomem odczuwanego stresu. Wskaźnik ten jest wyższy u mężczyzn. Wielkość efektu wskazuje, że prawdopodobieństwo uzyskania wyższej wartości w pierwszej grupie (wśród kobiet) wynosi 26%, a w drugiej (wśród mężczyzn) 74% ($CL = 0,26$).

3.2 Problem drugi - test istotności różnic dla danych niezależnych

Wykres:



Rysunek 2: Wykres drugi - rozkład współczynnika ilości snu względem płci

Hipotezy testowe:

H0: Mediana ilości snu u kobiet jest równa medianie ilości snu u mężczyzn.

H1: Mediana ilości snu u kobiet nie jest równa medianie ilości snu u mężczyzn.

Tabela 2: Porównanie wskaźnika ilości snu między płciami

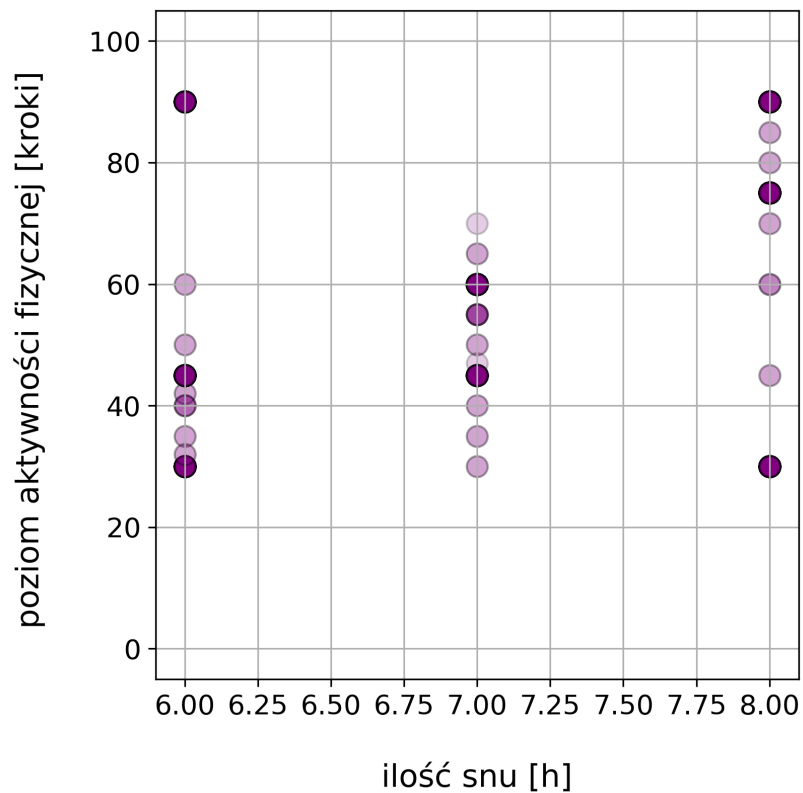
płeć	ilość snu			test Manna–Whitneya		
	n	med	IQR	U	p^*	CL
kobieta	185	7	2	18461	0,31	0,53
mężczyzna	189	7	2			

* – $p \leq 0,05$

Wniosek: Na poziomie istotności $p = 0,05$ nie możemy odrzucić hipotezy zerowej (H_0), mówiącej o równości mediany współczynnika ilości snu w obu analizowanych grupach. Nie jesteśmy w stanie stwierdzić czy kobiety ($med = 7$, $IQR = 2$) i mężczyźni ($med = 7$, $IQR = 2$) różnią się ilością snu. Wielkość efektu wskazuje, że prawdopodobieństwo uzyskania wyższej wartości w pierwszej grupie (wśród kobiet) wynosi 53%, a w drugiej (wśród mężczyzn) 47% ($CL = 0,53$).

3.3 Problem trzeci - test siły związku

Wykres:



Rysunek 3: Wykres trzeci - rozkład współczynnika poziomu aktywności fizycznej i ilości snu względem siebie

Hipotezy testowe:

H0: Korelacja między wskaźnikiem ilości snu a wskaźnikiem poziomu aktywności fizycznej wynosi 0.

H1: Korelacja między wskaźnikiem ilości snu a wskaźnikiem poziomu aktywności fizycznej jest różna od 0.

Tabela 3: Statystyki opisowe i współczynnik korelacji dla wskaźnika ilości snu i wskaźnika poziomu aktywności fizycznej

zmienna	rozkład zmiennej					test korelacji Spearmana	
	n	med	IQR	min	max	r_S	p
ilość snu	374	7	2	6	8	0,32	$\leq 0,001$
poziom aktywności fizycznej	374	59	30	30	90	*	

* – $p \leq 0,05$

Wniosek: Zgodnie z poziomem $p = 0,05$ odrzucamy hipotezę zerową (H_0), która mówi o korelacji równej zero między wskaźnikiem ilości snu a wskaźnikiem poziomu aktywności fizycznej. Między ilością snu a poziomem aktywności występuje korelacja ($r_S = 0,32$, $p \leq 0,001$), która jest dodatnia (im większa ilość snu tym większy poziom aktywności fizycznej). Wielkość efektu jest równa wielkości współczynnika korelacji ($r_S = 0,32$), co w tym przypadku oznacza słaby związek między zmiennymi.

3.4 Problem czwarty - test siły związku

Wykres:



Rysunek 4: Wykres czwarty - rozkład współczynnika jakości snu względem współczynnika poziomu stresu

Hipotezy testowe:

H0: Korelacja między wskaźnikiem poziomu stresu a wskaźnikiem jakości snu wynosi 0.

H1: Korelacja między wskaźnikiem poziomu stresu a wskaźnikiem jakości snu jest różna od 0.

Tabela 4: Statystyki opisowe i współczynnik korelacji dla wskaźnika poziomu stresu i wskaźnika jakości snu

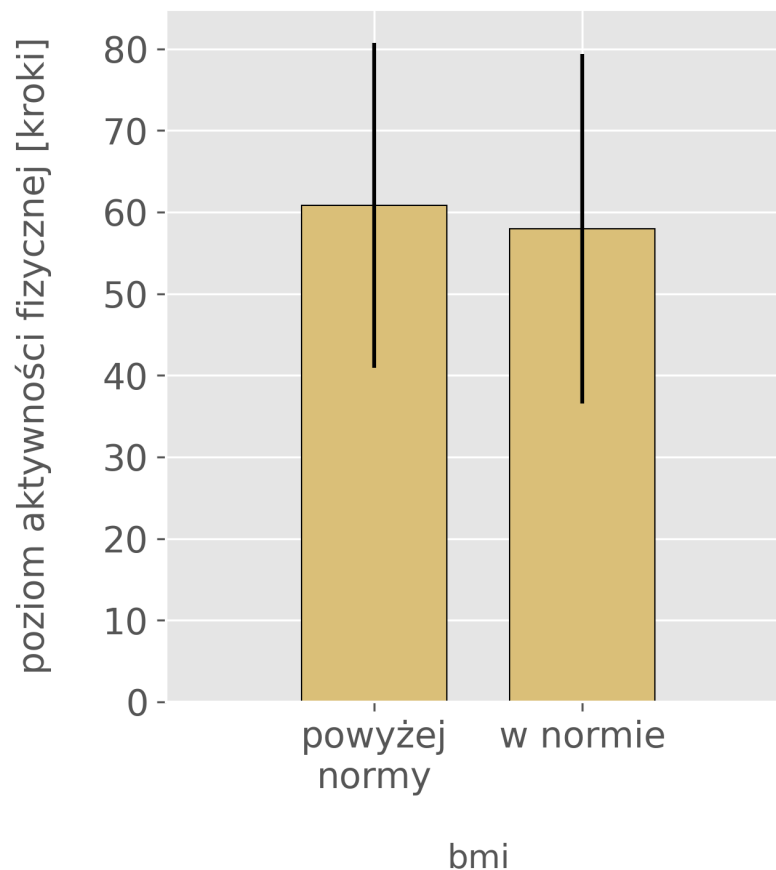
zmienna	rozkład zmiennej					test korelacji Spearmana	
	n	med	IQR	min	max	r_S	p
poziom stresu	374	5	2	3	8	-0,90	$\leq 0,001$
jakość snu	374	7	2	4	9		*

* – $p \leq 0,05$

Wniosek: Zgodnie z poziomem $p = 0,05$ odrzucamy hipotezę zerową (H_0), która mówi o korelacji równej zero między wskaźnikiem poziomu stresu a wskaźnikiem jakości snu. Między poziomem stresu a jakością snu występuje korelacja ($r_S = -0,90$, $p \leq 0,001$), która jest ujemna (im gorsza jakość snu tym większy poziom stresu). Wielkość efektu ($r_S = -0,90$) oznacza bardzo silny związek między zmiennymi.

3.5 Problem piąty - test istotności różnic dla danych niezależnych

Wykres:



Rysunek 5: Wykres piąty - rozkład współczynnika poziomu aktywności fizycznej względem bmi

Hipotezy testowe:

H0: Mediana poziomu aktywności fizycznej u osób z BMI w normie jest równa medianie poziomu aktywności fizycznej u osób ze wskaźnikiem BMI powyżej normy.

H1: Mediana poziomu aktywności fizycznej u osób z BMI w normie nie jest

równa medianie poziomu aktywności fizycznej u osób ze wskaźnikiem BMI powyżej normy.

Tabela 5: Porównanie wskaźnika poziomu aktywności fizycznej między osobami z różnym wskaźnikiem BMI

BMI	ilość snu			test Manna–Whitneya		
	n	med	IQR	U	p^*	CL
powyżej normy	158	61	30	18403	0,18	0,54
w normie	216	58	45			

* – $p \leq 0,05$

Wniosek: Na poziomie istotności $p = 0,05$ nie możemy odrzucić hipotezy zerowej (H_0), mówiącej o równości mediany współczynnika poziomu aktywności fizycznej w obu analizowanych grupach. Nie jesteśmy w stanie stwierdzić czy osoby ze wskaźnikiem BMI w normie ($med = 58$; $IQR = 45$) i powyżej normy ($med = 61$, $IQR = 30$) różnią się poziomem aktywności fizycznej. Wielkość efektu wskazuje, że prawdopodobieństwo uzyskania wyższej wartości w pierwszej grupie (wśród osób ze wskaźnikiem BMI powyżej normy) wynosi 54%, a w drugiej (wśród osób ze wskaźnikiem BMI w normie) 46% ($CL = 0,54$).